

An Approach to Data Prefetching Using 2-Dimensional Selection Criteria

Jean Sung¹Sebastian Krupa¹Andrew Fishberg¹Josef Spjut^{1,2}

Abstract—We propose an approach to data memory prefetching which augments the standard prefetch buffer with selection criteria based on performance and usage pattern of a given instruction. This approach is built on top of a pattern matching based prefetcher, specifically one which can choose between a stream, a stride, or a stream followed by a stride. We track the most recently called instructions to make a decision on the quantity of data to prefetch next. The decision is based on the frequency with which these instructions are called and the hit/miss rate of the prefetcher. In our approach, we separate the amount of data to prefetch into three categories: a high degree, a standard degree and a low degree. We ran tests on different values for the high prefetch degree, standard prefetch degree and low prefetch degree to determine that the most optimal combination was 1, 4, 8 lines respectively. The 2 dimensional selection criteria improved the performance of the prefetcher by up to 9.5% over the first data prefetching championship winner. Unfortunately performance also fell by as much as 14%, but remained similar on average across all of the benchmarks we tested.

I. INTRODUCTION

Prefetchers are a valuable asset to any modern performance motivated processor where the performance of the memory sub-system has a large impact on overall system performance [1]. Main memory accesses take hundreds of cycles while any on-chip memory access can be completed in a handful of cycles. Prefetching is powerful counter-measure that rarely hurts performance when implemented well [2]. Specifically, prefetching occurs when a device brings data or instructions from slower memory into the cache before they are explicitly requested or needed. To enable efficient preemptive fetching, a strong prefetching algorithm (i.e. an algorithm that frequently predicts what may be needed in advance) must be developed and utilized. For example, when a memory location is requested, a trivial prefetcher may also proactively grab the data that resides in the next memory address(es) [3]. Ideally, this would assist in tasks that require consecutive iterative memory access. However, for applications with other kinds of memory access patterns, this trivial prefetcher may reduce performance or cause unnecessary additional memory traffic and power consumption. A well performing prefetcher will only prefetch data that has a high probability of use.

Figure 1 shows a few example memory access patterns that we believe to be interesting and we use in the design of our prefetcher. Figure 1 (A) shows a standard streaming memory access pattern. A strided memory access pattern can be seen in Figure 1 (B) while Figure 1 (C) shows an access pattern that combines the two. We expect that most applications will

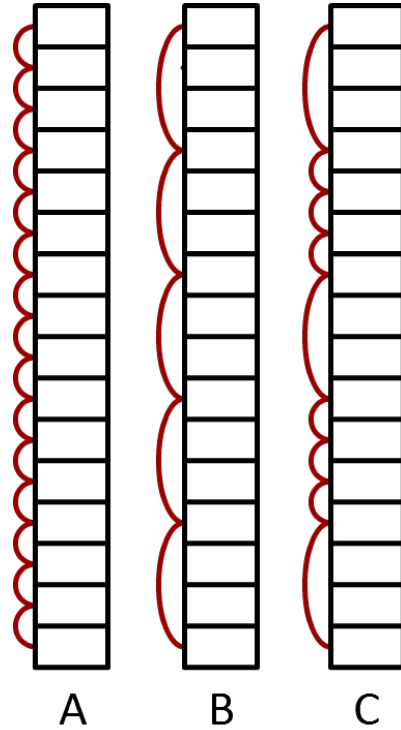


Fig. 1. Memory access patterns. (A) Stream (B) Stride (C) Stream + Stride

exhibit behavior that aligns with one of these patterns. Other applications do exist which access memory in a random pattern or according to strides or other patterns that we will fail to detect when only considering these three options.

II. PATTERN PREFETCHING

Our prefetcher is a simple stream and stride where the degree with which it prefetches instructions is influenced by two specific metrics. These metrics are stored on a per instruction basis where we store the 128 most recently called instructions. The architecture of our prefetcher can be seen in Figure 2. Each entry in our instruction table consists of a 64 bit instruction pointer, the last accessed data memory address (58 bits), the stride between the last strided access (6 bits), a 5 bit stream count, 32 bits for each to count misses and cycles, and a 6 bit lru value for replacing table entries totaling 203 bits per entry. Since we allow for a 128 entry table, a total of 21924 bytes of storage are required for our prefetcher. Please note that we introduce a performance optimization later in this section that increases the storage overhead slightly.

¹Harvey Mudd College, Claremont, CA

²NVIDIA Corp., Santa Clara, CA

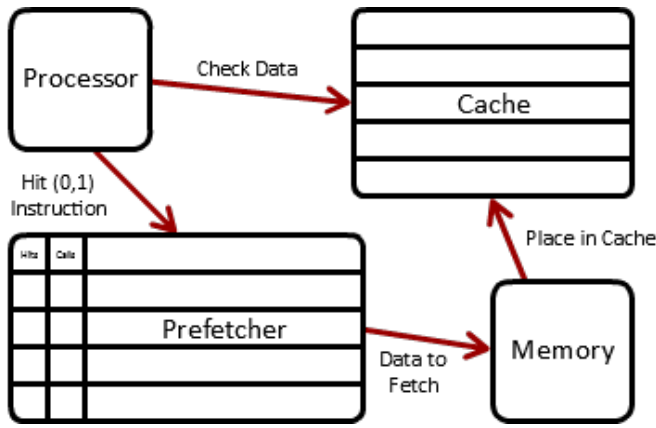


Fig. 2. Prefetcher Architecture

A. Usage Metric

The first metric we track is how commonly the instruction is called relative to the other instructions in the table. Each instruction is either flagged as "common" or "uncommon". We first sort all of the instructions by the number of times they have been referenced since being added to the tracker. The 78 instructions with the lowest reference count are marked as "uncommon" since they have been called infrequently relative to the other instructions.

The 50 most common instructions are considered common and the other 78 are considered uncommon. Newly tracked instructions should initially appear as uncommon until the prefetcher sees enough references to gain confidence, which is why we have fewer instructions marked as common. An alternate approach would be to tag new instructions as such before considering them in the common/uncommon sorting, but this would use space and add complexity. Furthermore, the current approach only makes it so that new instructions prefetch fewer lines of data until it is discovered if they are common or not. Since the prefetcher wants to flood the cache, a new tag would have a similar result.

B. Miss Metric

The other metric we use is how often the instruction missed the L2 cache. This allowed the prefetcher to know the miss rates for the instructions in our buffer. We can now judge how the prefetcher influences system performance for each instruction relative to the other instructions in the buffer. If an instruction performs poorly, prefetching more data may improve performance, whereas if it performs well standard prefetching should suffice.

To decide how to quantify which instructions are doing well and which are doing poorly, we use a threshold parameter. If the threshold is 50%, for an instruction to be considered to be doing well, it has to have a miss rate lower than 64 instructions currently being tracked. However, this is problematic since new instructions will pad the data with extreme values, i.e. an instruction called once that hit would be considered to be doing really well and one that missed would be considered to be doing really poorly.

usage	low miss rate	high miss rate
common	standard	high
uncommon	low	standard

TABLE I
PREFETCHING DEGREE

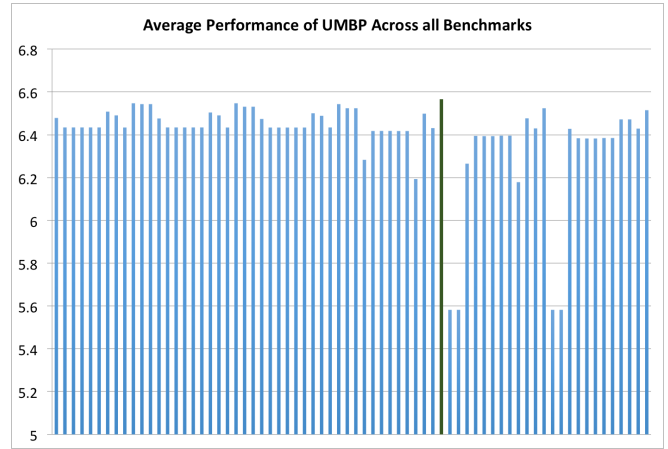


Fig. 3. Graph of varying parameters used to determine optimal values for the percentage threshold as well as the three prefetching degrees. The x axis represents the variation of these parameters and the green bar represents the optimal combination of parameters.

Instead we compare the miss rate of the current instruction to the miss rates of the 50 most common instructions along with 20 random less common instructions. We don't want to just use the most common instructions since those can have abnormally high or low miss rates, so we seed in less common instructions to alleviate this problem. This 70 entry table requires 64 bits per entry, which combines with the first required table for a total of 32768 (2^{15}) bits of storage.

C. Usage and Miss-based Prefetching

Combining the 2 metrics we get a matrix as seen in Figure I. We chose to have 3 discrete prefetching degrees, a standard degree which most instructions will use, a high degree which only instructions that are common but performing poorly will use, and a low degree which uncommon instructions that are performing well will use. This should prioritize prefetching for instructions that are affecting performance most.

We ran a set of tests using the simulation infrastructure described in Section III to determine the optimal values for the three prefetching degrees as well as the threshold percentage. We tested using the provided benchmark traces as well as a set of additional benchmark traces. The optimal miss rate threshold was around 25-40 % for all combinations of prefetching degrees. The optimal number as seen in Figure 3 of lines to prefetch were found to be 1 for the low degree, 4 for the standard degree, and 8 for the high degree.

We found that the low degree should be as low as possible always, but never zero. This means that the benefits from a single stream or a single stride almost always outweigh the

Level	Size	Associativity	Latency
L1 Data	16KB	8-way	4 cycles
L2 Data	128KB	8-way	10 cycles
L3 Data	Varies	16-way	20 cycles
Instruction	N/A	N/A	N/A

TABLE II
SIMULATION CACHE CONFIGURATION

costs of prefetching an extra line of data. However, since in these case, these were instructions that were performing well, no further prefetching was necessary, thus increasing this value would only flood the cache.

The standard and high degrees both benefited from prefetching a large number of cache blocks. Increasing the standard degree up to a certain point had a slight positive increase in performance whereas increasing the high degree had a considerable increase in performance. This corroborates our hypothesis since we suspected that the common instructions that are doing poorly needed the largest amount of prefetched data.

III. RESULTS

In order to test our prefetcher implementation, we build a prefetcher using the simulator developed by Pugsley [4] for the Second Data Prefetching Championship. The simulator is a trace-based simulator that models a simple out-of-order core with a 256-entry instruction window. The simulated processor has 6-wide superscalar issue and assumes perfect branch prediction. The cache hierarchy configuration is found in Table II. Instruction caches are not modeled and the L3 Data cache size varies with simulation. Data caches are simulated as inclusive and the rest of the simulation parameters can be found on the data prefetching championship website [4]. Note that the version of the simulator used to generate all of our results is the March 18th version that has a known bug that prevents prefetches from being promoted to demands at the memory controller.

In order to compare our prefetcher to other styles of prefetching, we used the set of prefetchers included with the second data prefetching championship source code [4]. The first comparison point is called *skeleton*, which represents the absence of a prefetcher, or a prefetcher that does nothing. The second prefetcher, called *next line prefetcher* simply fetches the next line whenever an address is requested. Neither of these prefetchers requires any memory table or storage.

The *stream prefetcher* works off of spatial locality by establishing a "stream" in which the program first checks if the accessed page is one of the 64 pages being kept track of. If it isn't, the oldest page is overwritten with this new page. It then checks to see if the page offset has increased or decreased. If it has increased, it checks to see which direction was already stored. If the direction of the page offset and the direction stored are the same, the confidence increases, otherwise the confidence is reset to 0 and the direction stored is flipped. It uses 1560 bytes of storage to keep track of these

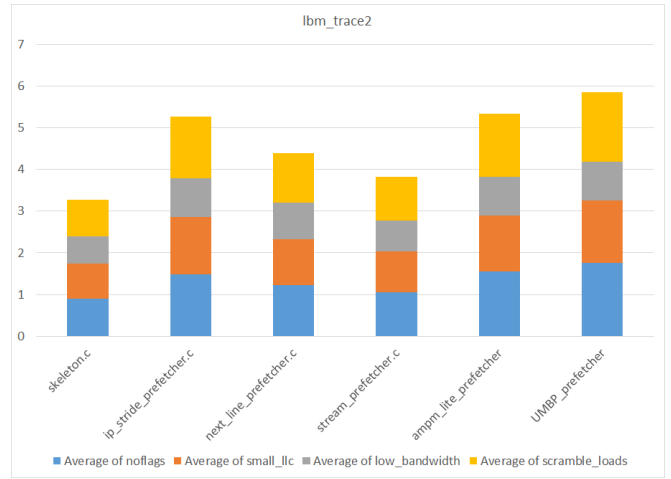


Fig. 4. Best Case Scenario for Usage and Miss based Prefetcher

pages. If the confidence is high enough (2 or greater), then the next two lines are prefetched into either the L2 cache or the LLC depending on occupancy.

The instruction pointer based stride prefetcher, called *ip stride prefetcher* in the tables, keeps track of which addresses a specific instruction accessed most recently. It then assumes that future addresses to be accessed will be some fixed stride length away. The inner workings are similar to the stream, but this prefetcher stores and checks instruction pointers rather than pages. If it has seen a specific stride at between the last 3 memory addresses accessed, it prefetches the next 3 addresses that are that stride away. It requires 32780 bytes of storage for the information it tracks.

The final point of comparison is Access Map Pattern Matching, called *ampm lite* in the tables, is a simplified implementation of the winning data prefetcher from the first data prefetching championship [5]. We believe this is the most competitive point of comparison and will represent a high water mark for our comparison.

On average, the performance of the Usage and Miss-based prefetcher described in this paper was comparable to AMPM-lite. While we cannot include detailed results for every benchmark and configuration in this paper, more of the results can be found on github [6].

The prefetcher performed the best on the lbn trace, where it achieved an improvement of 9.5% over the AMPM lite prefetcher 4. The prefetcher performed poorly on gcc, where it was 14% lower than AMPM lite 5. In this case, it still performed better than the other provided prefetchers. On most other traces, the differences were small. Furthermore, the better performing traces such as GemsFDTD had negligible differences between our prefetcher and AMPM lite. The worse performing prefetchers tended to have more variability. These results imply that our prefetcher manages to handle some extreme cases which degrade performance not covered by the other prefetchers. Likewise, AMPM manages to catch some such cases that our prefetcher misses.

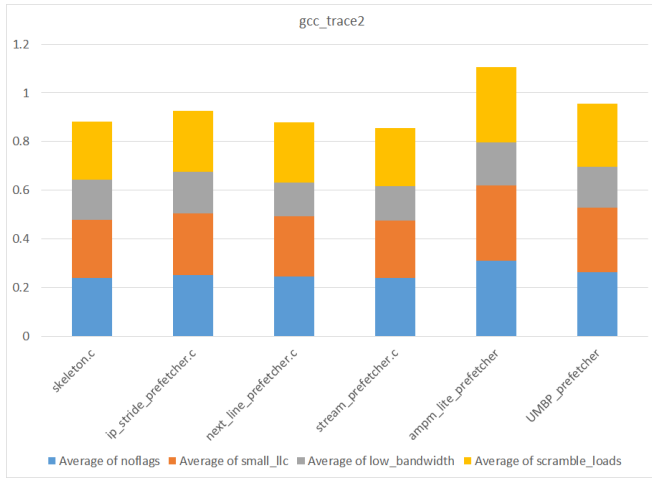


Fig. 5. Worse Case Scenario for Usage and Miss based Prefetcher

While we only present a selection of benchmark results in this paper, a larger set of simulations were run and the results can be found in the github repository [6]. We believe that making our source code available will allow anyone a desire to reproduce our results to do so. Furthermore, it should be straightforward to add functionality and produce new designs from our prefetcher.

IV. RELATED WORK

The huge CPU performance gains obtained through predictive prefetching has made it an active and worthwhile research field for several decades. Over the years, many increasingly sophisticated prefetching models have been suggested (i.e. sequential prefetcher, tagged sequential prefetcher, etc.) The most trivial prefetcher, the sequential prefetcher outlined within the Introduction, simply fetches the next line after accessing the cache [3].

The top three prefetchers submitted to the First Data Prefetching Championship (DPC-1) [7], titled AMPM (Access Map Pattern Matching), GHB-LDB (Global History Buffer - Local Delta Buffer), and PDFCM (Prefetching base on a Differential Finite Context Machine), utilized significantly more sophisticated and novel approaches to the prefetching problem.

- The AMPM prefetcher [5] divided memory into hot zones that were tracked using a 2-bit vectors to locate the stride zone. Our prefetcher instead tracks the last stride with a 6-bit integer.
- The GHB-LDB prefetcher [8] improved upon existing prefetcher models by including global correlation and common stride prefetching. Our prefetcher in contrast simply tracks variable strides and streams.
- The PDFCM prefetcher [9] utilized hash-based History Tables and Delta Tables to prefetch arbitrary degrees and distances. Our prefetcher utilizes a simpler direct look-up solution that makes prefetching decisions based on the frequency of instruction usage and misses.

V. CONCLUSION

Here we have presented a simple yet effective prefetching solution that utilizes the frequency of instruction usage and the miss rate to inform prefetching decisions. Specifically, we tune the amount of data we prefetch based on these criteria by having three prefetching degrees. Most instructions prefetch a standard amount of data, the least common instructions that have low miss rates prefetch fewer lines of data, and finally, instructions that miss a lot and are common prefetch more lines of data.

Our prefetching solution was tested using benchmarks provided by the Second Data Prefetching Championship (DPC-2). We found that the UMBP prefetcher on average performed similarly to the provided AMPM prefetcher; however, it did perform better on traces that had lower overall IPC values. Other benchmarks showed flat performance or even some loss of performance. We believe that the approach presented here could result in more performance if more effort is spent on tuning the prefetcher.

ACKNOWLEDGMENT

This work would not have been possible without the simulator written by Seth Pugsley and the support provided on the Second Data Prefetching Championship mailing list.

REFERENCES

- [1] N. P. Jouppi, "Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers," in *Computer Architecture, 1990. Proceedings., 17th Annual International Symposium on*, pp. 364–373, IEEE, 1990.
- [2] S. Cepeda, "What you need to know about prefetching," 2009. <https://software.intel.com/en-us/blogs/2009/08/24/what-you-need-to-know-about-prefetching>.
- [3] A. J. Smith, "Cache memories," *ACM Computing Surveys (CSUR)*, vol. 14, no. 3, pp. 473–530, 1982.
- [4] S. H. Pugsley, "The second data prefetching championship," 2015. <http://comparch-conf.gatech.edu/dpc2/>.
- [5] Y. Ishii, M. Inaba, and K. Hiraki, "Access map pattern matching for high performance data cache prefetch," *Journal of Instruction-Level Parallelism*, vol. 13, pp. 1–24, 2011.
- [6] J. Sung, S. Krupa, A. Fishbery, and J. Spjut, "Charlab dpc2 github repository," 2015. <https://github.com/charlab/dpc2>.
- [7] M. Grannaes, M. Jahre, and L. Natvig, "Multi-level hardware prefetching using low complexity delta correlating prediction tables with partial matching," in *High Performance Embedded Architectures and Compilers* (Y. N. Patt, P. Foglia, E. Duesterwald, P. Faraboschi, and X. Martorell, eds.), vol. 5952 of *Lecture Notes in Computer Science*, pp. 247–261, Springer Berlin Heidelberg, 2010.
- [8] M. Dimitrov and H. Zhou, "Combining local and global history for high performance data prefetching," *Journal of Instruction-Level Parallelism*, vol. 13, pp. 1–14, 2011.
- [9] L. M. Ramos, J. L. Briz, P. E. Ibáñez, and V. Viñals, "Multi-level adaptive prefetching based on performance gradient tracking," *Journal of Instruction-Level Parallelism*, vol. 13, pp. 1–14, 2011.